# Evolutionary Origins of Transcription Factor Binding Site Clusters

Xin He,†,[1] Thyago S.P.C. Duque,†,[2] and Saurabh Sinha*,[2]

[1]Department of Biochemistry, University of California at San Francisco

[2]Department of Computer Science, University of Illinois at Urbana-Champaign

†These authors contributed equally to this work.

***Corresponding author:** E-mail: sinhas@illinois.edu.

**Associate editor:** Sudhir Kumar

## Abstract

Empirical studies have revealed that regulatory DNA sequences such as enhancers or promoters often harbor multiple binding sites for the same transcription factor. Such "homotypic site clustering" has been hypothesized as arising out of functional requirements of the sequences. Here, we propose an alternative explanation of this phenomenon that multisite enhancers are common because they are favored by evolutionary sampling of the genotype–phenotype landscape. To test this hypothesis, we developed a new computational framework specialized for population genetic simulations of enhancer evolution. It uses a thermodynamics-based model of enhancer function, integrating information from strong as well as weak binding sites, to determine the strength of selection. Using this framework, we found that even when simpler genotypes exist for a desired strength of regulation, relatively complex genotypes (enhancers with more sites) are more readily reached by the simulated evolutionary process. We show that there are more ways to "build" a fit genotype with many weak sites than with a few strong sites, and this is why evolution finds complex genotypes more often. Our claims are consistent with an empirical analysis of binding site content in enhancers characterized in *Drosophila melanogaster* and their orthologs in other *Drosophila* species. We also characterized a subtle but significant difference between genotypes likely to be sampled by evolution and equally fit genotypes one would obtain by uniform sampling of the fitness landscape, that is, an "evolutionary signature" in enhancer sequences. Finally, we investigated potential effects of other factors, such as rugged fitness landscapes, short local duplications, and noise characteristics of enhancers, on the emergence of homotypic site clustering.

Homotypic site clustering is an important contributor to the complexity and function of *cis*-regulatory sequences. This work provides a simple null hypothesis for its origin, against which alternative adaptationist explanations may be evaluated, and cautions against "evolutionary mirages" present in common features of genomic sequence. The quantitative framework we develop here can be used more generally to understand how mechanisms of enhancer action influence their composition and evolution.

**Key words:** enhancer evolution, homotypic site clustering, complex genotypes, thermodynamic model.

Research article

## Introduction

Enhancers involved in metazoan development have been known to harbor multiple binding sites for the same transcription factor (TF), a phenomenon known as "homotypic clustering" (HTC). This has been documented in invertebrate (Berman et al. 2002; Markstein et al. 2002; Li et al. 2007) and vertebrate (Sinha et al. 2008; Gotea et al. 2010) genomes alike and is the basis for several genome-wide enhancer prediction tools (Berman et al. 2002; Markstein et al. 2002; Lifanov et al. 2003; Sinha et al. 2008; Gotea et al. 2010). Several explanations have been offered for this common empirical observation. The common explanation is that multiple homotypic sites in an enhancer (or promoter) are required for the enhancer's transcriptional efficacy, the desired gene expression levels and ultimately for organismal fitness (Sauer et al. 1995; Hertel et al. 1997). That is, the observed site multiplicity is ostensibly due to selective forces (e.g., Shultzaberger et al. 2010). For example, various theories have proposed that site clusters may 1) facilitate lateral diffusion of TF molecules along the DNA, thereby increasing the effective protein concentration (Kim et al. 1987; Coleman and Pugh 1995) or 2) increase occupancy nonlinearly through cooperative interactions among sites (Giniger and Ptashne 1988; Hertel et al. 1997) or through simultaneous interaction with the basal transcriptional machinery (BTM) (Lin et al. 1990; Anderson and Freytag 1991; He et al. 2010). Indeed, nonlinear transcriptional response to protein concentration is believed to be important for various phenotypes (Porcher and Dostatni 2010), again suggesting that HTC may be common due to a selective advantage.

However, common features observed in a class of genomic elements may not be due to functional constraints alone; they may also result from properties of the "fitness landscape" (Mustonen et al. 2008) and from evolutionary sampling of this landscape (Lusk and Eisen 2010). (The space of all possible nucleotide sequences, i.e., genotypes, with a fitness value assigned to every genotype, is henceforth called

the fitness landscape.) We hypothesized that the fitness landscape and its evolutionary sampling play an important role in the origin of HTC. For instance, a "simple" sequence with one or two perfect binding sites and a "complex" sequence with a number of weaker sites may be equally effective at activating a gene, but complex sequences may be far more abundant and thus favored by evolution. Here, we explore the evolutionary origins of homotypic site clusters in enhancers, through direct examination of the fitness landscape and by simulating the evolution of a simple enhancer. We find that evolution favors complex genotypes even when simpler (more parsimonious) genotypes of comparable fitness exist. This is largely because the space of fit genotypes has more of the former than the latter. Our findings are consistent with an empirical analysis of binding site multiplicities in experimentally characterized enhancers in *Drosophila melanogaster*.

Our results caution against "evolutionary mirages" (Lusk and Eisen 2010), where properties of the evolutionary process lead to genotypic properties that may appear to have mechanistic origins. In particular, they suggest an evolutionary "null hypothesis" for the phenomenon of HTC, against which alternative explanations, mechanistic or evolutionary, may be assessed in the future.

## Materials and Methods

### Quantitative Model of Enhancer Function

We use the GEMSTAT software (He et al. 2010) to predict the expression driven by any enhancer sequence, given the binding specificity and concentration of the TF. The transcriptional output of an enhancer is assumed to be proportional to the probability that the BTM occupies the gene's promoter. The model enumerates every possible configuration of TF molecules and the BTM bound to their respective binding sites. Thus, for example, for a sequence with two possible binding sites, eight different configurations would be considered (fig. 1C), four "OFF" configurations (BTM is not occupied) and four "ON" configurations. Any configuration is associated with a statistical weight, which is the probability of that configuration and is determined by energetic contributions from all molecular interactions included in that configuration. Specifically, the weight of an OFF configuration is determined by interactions between TF molecules and their binding sites (fig. 1C, left panel) and the weight of an ON configuration is determined by TF-BTM interactions (fig. 1C, right panel) in addition to the TF-site interactions. The fractional occupancy of the BTM, and thus the gene expression level, is given by the total statistical weight of all ON configurations (eq. 1 in He et al. 2010), relative to that of all configurations. The expression profile driven by an enhancer is predicted by repeating the above process of expression prediction for every distinct value of TF concentration along the axis. The GEMSTAT model has parameters related to two kinds of cooperativity: 1) DNA-binding cooperativity: interactions between adjacent occupied binding sites and 2) transcriptional synergy: the synergistic effects of multiple binding sites simultaneously recruiting the transcriptional machinery, controlled by the $N_{MA}$ parameter. All our experiments reported in the text were performed under the setting of no cooperative DNA binding and a modest level of transcriptional synergy ($N_{MA} = 2$, i.e., only two bound sites could simultaneously activate transcription). We explored different values of $N_{MA}$, as reported in supplementary figures S3 and S12 (Supplementary Material online).

### Strength of Binding Sites and TF Occupancy

The strength of a binding site is defined as its binding affinity relative to the strongest ("consensus") site. It is a number between 0 and 1, and a site with a relative affinity of 0.1, for example, is 10 times weaker than the optimal site, in terms of association constant. Let LLR($s$) be the log likelihood ratio score of site $s$, computed based on the known position weight matrix (PWM) of the TF and the background nucleotide distribution (Stormo 2000). The site's strength (relative affinity) is computed as $\exp(\mathrm{LLR}(s) - \mathrm{LLR}(s_{opt}))$, where $s_{opt}$ is the optimal site.

The TF occupancy at an enhancer is defined as the sum of the fractional occupancy of all sites in the enhancer, at maximum TF concentration, as computed by the GEMSTAT model. Fractional occupancy of a site is given by the total statistical weight of all configurations where the site is bound relative to that of all configurations.

### Analytical Estimation of Number of Genotypes with $k$ Sites

The relative affinity threshold is converted to a $P$ value $p$ of the site LLR, and the (relative) number of genotypes with at least $k$ sites at this threshold is computed as $\binom{L}{k}2^k p^k$, where $L$ is the length of the enhancer sequence. Taking differences between successive values of $k$ gives the desired number of genotypes, up to a constant of proportionality.

### Fitness Functional

Given two expression profiles $u$ (real) and $v$ (predicted), as $n$-dimensional vectors of expression levels, a natural way to define the fitness functional $F$ would be to use a Euclidian distance between $u$ and $v$. Our evolutionary simulation framework is meant to be generally usable with any "real" expression profile (not just the one shown in fig. 1A), and our previous work (Kazemian et al. 2010) showed a particular heuristic score called the "pattern generating potential" (PGP) to have the most desirable properties for this purpose. Here, we defined $F$ based on a PGP-like approach but unlike PGP, we allowed $u$ to be continuous profile, as follows:

1. compute reward $= \sum u_i \min(u_i, v_i) / \sum u_i^2$,
2. compute penalty $= \sum (u_{max} - u_i) \max(0, v_i - u_i) / \sum (u_{max} - u_i)^2$,
3. obtain $F = [\max(0, \text{reward} - \text{penalty})]^2$.

The reward term is sensitive to the underprediction of expression levels where true expression is high, whereas the penalty term is sensitive to overprediction of expression levels where true expression is low.

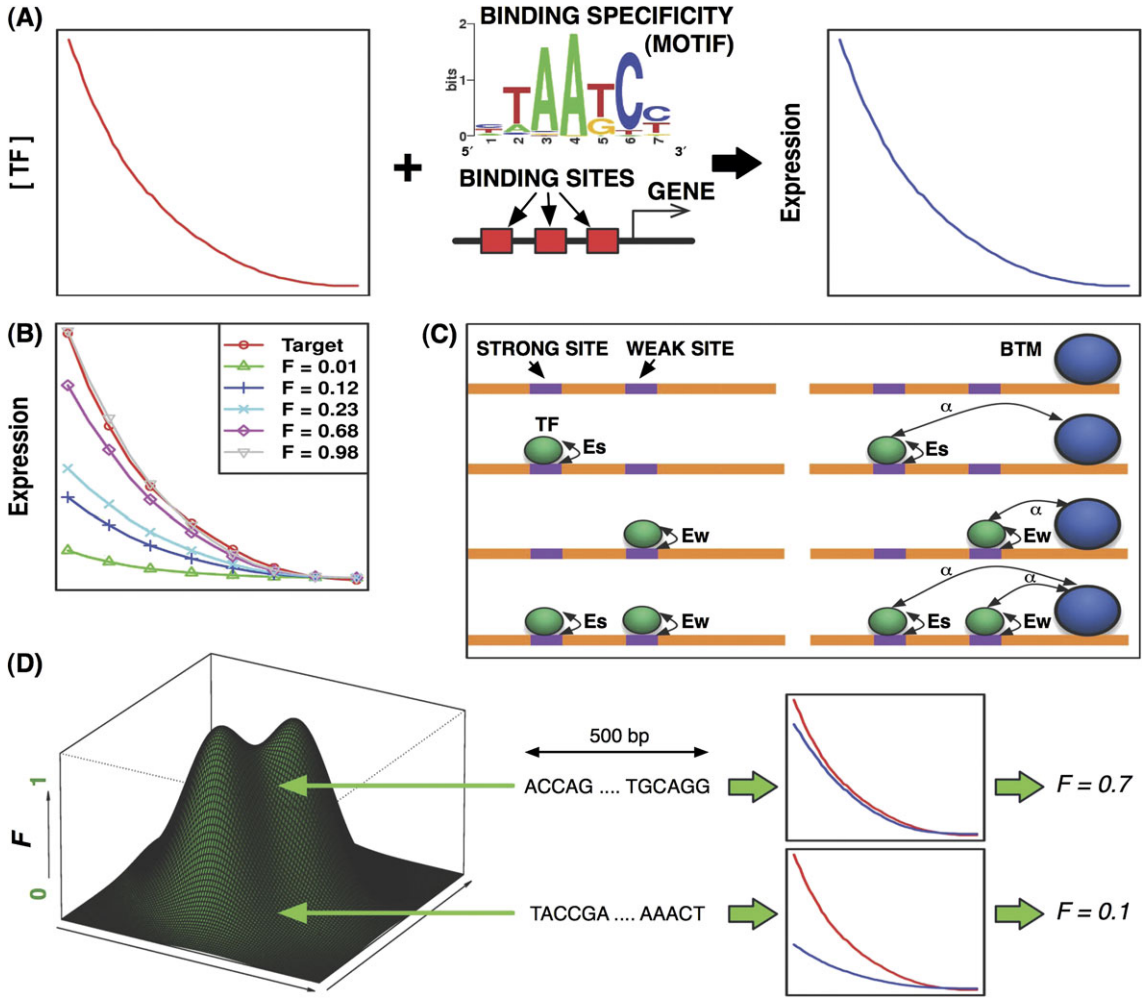


**Fig. 1.** A model system for studying evolution of enhancers. (*A*) The spatial expression pattern of TF (left panel, TF concentration plotted along the anterior/posterior axis) is read by an enhancer (middle, bottom) with sites matching the TF motif (middle, top), and the result is a spatial expression pattern of the gene regulated by the enhancer (right panel). (*B*) Example gene expression profiles compared with the target profile (red) and associated fitness functional (*F*) values. (*C*) A thermodynamic model of enhancer function. Shown is a sequence with two binding sites (one strong, one weak), which may exist in eight possible configurations of TF molecules (green circles) bound to these sites, and in four of which the BTM is bound to the promoter. The terms $E_s$ and $E_w$ represent the energetic interactions between a TF molecule and its site, and arrows labeled $\alpha$ denote interactions between TF molecules and BTM. Transcription is assumed to be initiated only when BTM is bound; thus, the total probability of the four configurations on the right determines the activation level of the gene due to this enhancer (for details, see Materials and Methods). (*D*) A cartoon illustration of the fitness landscape. All possible sequences are points on the horizontal plane. Each sequence corresponds to an expression pattern, which determines its fitness functional (*F*) value.

It is worth noting that the exponentially decaying expression pattern of the TF and the regulated gene are not directly relevant to our claims; rather, we want to examine the simple situation when the gene expression level follows the TF concentration levels in a linear fashion. In fact, if we simplify the fitness functional substantially to have only two levels of TF concentration (high and low) and demand the gene expression to respond linearly to the TF level, our conclusions continue to hold (supplementary fig. S11, Supplementary Material online).

## Evolutionary Simulation

We simulate the evolution of a fixed-size population of sequences, following the Wright–Fisher process. At each generation, random mutations are introduced into the population. For simplicity, we consider only point mutations, though the effect of indels was studied in a separate experiment (next section). The fitness value of each sequence is calculated as $1 + s$, where $s$ is the selection coefficient, related to the fitness functional ($F$), we defined earlier as $s = FK$, where $K$ is a parameter. The probability of a sequence being sampled for inclusion in the next generation is proportional to $(1 + s)$. Thus, an individual with $F = 1$ is expected to produce $(1 + K)$ times more offspring than an individual with $F = 0$. We note that the simulation must recompute the expression profile for every mutant individual in each generation and thus relies upon efficient implementation of the GEMSTAT model.

## Parameterization of the Expression Model and Evolutionary Simulation

The two main parameters of the GEMSTAT model are the "DNA-binding" parameter ($\beta$) and the "activation strength" parameter ($\alpha$). Our default parameter settings were $\beta = 5$,

$\alpha = 2$. These values were obtained from a separate exercise where we simultaneously modeled the expression profiles of 20 A/P axis patterning enhancers (and the nonexpression of equally many random sequences), using the binding specificities (motifs) of six different TFs (see supplementary fig. S4, Supplementary Material online). The *Bicoid* TF, whose motif and concentration profile we have used throughout our study, was assigned the above values (approximately) in the trained model. To get some intuition into what these values mean, we note that $\beta = 5$ implies that the consensus binding site for the TF has a fractional occupancy of 5/6 at maximum TF concentration. Likewise, $\alpha = 2$ implies that a site with fractional occupancy $\approx 1$ induces 2-fold activation of gene expression, and under our settings for synergistic activation, about ~7 high occupancy sites are needed to achieve 100-fold activation. This number is roughly consistent with the number of *Bicoid* sites found in the well-studied hunchback promoter that drives anterior expression. Furthermore, a simple calculation shows that with these parameter settings, a random sequence of length 500 bp is expected to show no expression. Thus, we believe that our default settings for the thermodynamic model parameters are realistic. We also repeated our simulations with an alternative setting ($\beta = 1$, $\alpha = 5$, supplementary fig. S5, Supplementary Material online) and found little difference in the main observations reported above.

The key parameters in the evolutionary simulations are the population size ($2N$), the mutation rate per nucleotide per generation ($\mu$) (or equivalently, $2N\mu$), and the selection coefficient ($s$) (or equivalently, $4Ns$). Default settings of these parameters were $2N = 100$, $2N\mu = 10^{-3}$ and $4Ns = 100$. Standard values of the population size and mutation rate, from the literature, are $2N \sim 10^5$–$10^6$ (Thornton and Andolfatto 2006) and $\mu \sim 10^{-8}$–$10^{-9}$ (Drake et al. 1998) giving us $2N\mu$ in the range of 0.01–0.0001, which is approximately what we set it to be. We used time rescaling (Hoggart et al. 2007) to speed up our simulations. Here, the population size is scaled down by a constant (we used $\lambda = 1000$), keeping $2N\mu$ and $4Ns$ unchanged; $t$ generations of simulation in this scheme is approximately equivalent to $\lambda t$ generations of simulation in the absence of rescaling. Thus, the default setting of $2N = 100$ is equivalent to $2N = 10^5$ without time scaling. We repeated the simulations with a larger population size of $2N = 1000$ (equivalent to $2N = 10^6$, unscaled) and noted that the observed trends were unchanged (supplementary fig. S6, Supplementary Material online). We set the selection coefficient $s$ of a genotype as $s = FK$, where $F \in [0,1]$ is the fitness functional of the genotype and $K$ is the selection coefficient of the fittest genotype ($F = 1$) in relation to the least fit genotype ($F = 0$). The latter was set to a value of $50 \cdot 1/2N$ by default, indicating strong selection ($2NK = 50$). Note that in any one generation, there is a relatively small difference in $F$ between the fittest genotype and the wild type; this means that the effective selection coefficient $s$ for the fittest genotype is typically much smaller than $50/2N$. We also repeated our simulations with $2NK$ set to 10 and 20. Adaptation was often not observed in the sampled time at the former value, hence, the corresponding results are not shown. Results of simulations with $2NK = 20$ are shown in supplementary figure S7 (Supplementary Material online) and support our claims above. All simulations were performed in the absence of insertions and deletions, which have been suggested as important influences in the evolutionary dynamics of regulatory sequences (Sinha and Siggia 2005; Lusk and Eisen 2010). Although a detailed examination of this influence was not pursued here, we repeated our simulations with indels (at rates proposed in the literature) and found our observations about distributions of site multiplicity to be unchanged (supplementary fig. S8, Supplementary Material online).

### *Drosophila* Enhancers

We collected 21 *Bicoid*-driven enhancers from *D. melanogaster* with functions in anterior/posterior patterning (Ochoa-Espinosa et al. 2005; Halfon et al. 2008). Orthologous sequences from five other species in the melanogaster group (*D. ananassae*, *D. pseudoobscura*, *D. virilis*, *D. mojavensis*, and *D. grimshawi*) were extracted using the liftover tool http://genome.ucsc.edu/cgi-bin/hgLiftOver.

## Results

### Fitness Landscape and Evolutionary Simulations

We begin with the gene expression pattern that constitutes the phenotype for this study. Our enhancers will harbor binding sites for a single TF, whose concentration has an exponentially decaying pattern along an axis (fig. 1A). This mimics the concentration gradient of the morphogen *Bicoid* along the anterior/posterior (A/P) axis of the blastoderm-stage *Drosophila* embryo, but more generally, it reflects the fact that most TFs have spatial/temporal variability in their concentration. The binding specificity of the TF is assumed to be described by the *Bicoid* PWM (fig. 1A; Bergman et al. 2005). The expression pattern that a functional enhancer is required to encode is chosen to be identical in shape to the TF pattern, with 100-fold activation at the highest levels of the TF. (Note that to implement such a linear "readout" of the TF concentration gradient, an enhancer does not require cooperative interactions among its binding sites.) We next define a "fitness functional" (denoted by $F$) for any enhancer as a measure of how similar its induced expression profile is to the required profile. This is a number between 0 and 1 (with 1 indicating identity) and is derived from the PGP score in Kazemian et al. (2010) (fig. 1B and Materials and Methods).

An important component of our framework is the quantitative model that maps the enhancer sequence, along with the TF concentration and binding specificity, to gene expression. We use a model based on statistical thermodynamics that is very similar to that proposed by Shea and Ackers (1985) and discussed and refined by several recent studies (Buchler et al. 2003; Gertz et al. 2009; He et al. 2010). This model has been demonstrated to explain well the spatial patterning of early developmental genes in *Drosophila*

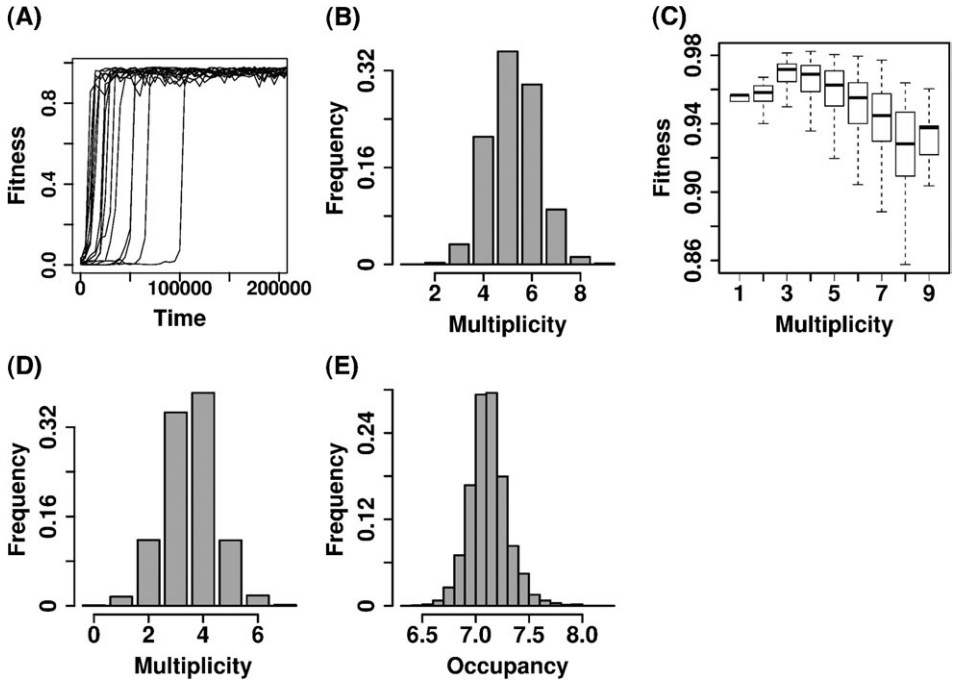


**Fig. 2.** Results of evolutionary simulations. (*A*) Time series of (average) population fitness, showing adaptation. Each curve represents the history of one population (truncated at 200,000 generations). (*B*) The distribution of site multiplicity (number of binding sites) in postadaptation genotypes. (Five random individuals with $F \geq 0.8$ were sampled from the population every 5000 generations.) The *x* axis is the number of sites (at relative affinity $\geq 0.25$) and the *y* axis is the frequency of genotypes with that multiplicity. (*C*) Box plot of fitness (*F*) values of genotypes with different site multiplicities. (*D*) Same as plot (*B*) but for a higher affinity threshold, 0.50. (*E*) Distribution of TF occupancy in postadaptation genotypes.

(Zinzen et al. 2006; Segal et al. 2008; He et al. 2010). A brief description of the model is provided in Materials and Methods (also fig. 1*C*), whereas details can be found in our earlier work (He et al. 2010). Importantly, even weak binding sites contribute to regulation in this model, thus allowing a prediction of the readout encoded by any sequence in the genotype space (and not just those with one or more sites above some threshold). Also, cooperative DNA binding of multiple TF molecules was excluded from the model, for reasons given later (Discussion). We examined the space of 500 bp long sequences (genotypes), their respective expression patterns (phenotypes) and the fitness functional values computed from the phenotypes (fig. 1*D*).

To characterize the distribution of fit genotypes that an evolutionary process would encounter, we performed Wright–Fisher simulations of a fixed-size population, where each individual is an enhancer genotype. Repeated rounds of random mutation and natural selection were applied to the evolving population, where strength of selection depends on the phenotype (expression pattern) of a sequence and its fitness. (For details and justification of evolutionary and biophysical parameters, see Materials and Methods.)

## The Space of Fit Genotypes Sampled by Evolution Shows a Relative Abundance of Complex Genotypes

Fifty independent evolutionary simulations were run for $10^6$ generations each; adaptation typically happened within $10^5$ generations, and the average fitness functional for the population stayed above $F = 0.8$ thereafter (fig. 2*A*). We sampled postadaptation genotypes from all simulations and examined the site multiplicity of this evolutionary sample of fit genotypes (fig. 2*B*). (Site multiplicity is the number of binding sites in the genotype, defined by a threshold on their binding affinity relative to that of the optimal site. The threshold used here is 0.25 times the binding affinity of a perfect site; for details, see Materials and Methods). While the most parsimonious genotypes sampled use only 1 above-threshold site, the mode of the distribution is at five sites, clearly demonstrating that evolutionary sampling favors genotypes with relatively high site multiplicity. This observed bias is not due to the complex genotypes in the pool having higher fitness (fig. 2*C*). At the same time, very complex genotypes (e.g., those with >7 sites at the threshold) are also rare in the evolutionary sample. The trends of figure 2*B* are also seen when defining sites with a stricter threshold of relative affinity $\geq 0.5$ (fig. 2*D*). We also plotted the genotype frequency at different values of the "occupancy" of the TF on the entire enhancer (fig. 2*E*). (Occupancy is defined by the thermodynamic model as the average number of sites bound by TF molecules and is independent of any threshold on site affinity; see Materials and Methods.) Clearly, the range of observed occupancy values is much smaller than the ranges of site multiplicity (fig. 2*B* and *D*). In other words, selection ensures that the genotypes sampled after adaptation lie in a narrow range of occupancy (which is closely related to fitness); however, the same occupancy level (and hence fitness) can be achieved through a wide range of site multiplicities.

## Causes of the Evolutionary Bias toward Complex Genotypes

### *Abundance of Complex Genotypes in the Fitness Landscape*

The distribution of genotypes sampled by evolution is shaped, to a large extent, by the fitness landscape. (For an expression for the equilibrium probability of sampling a genotype, as a function of its fitness functional $F$, e.g., see Sella and Hirsh 2005.) Thus, a possible explanation for the complex genotype bias seen above is that the fitness landscape has a relative abundance of such genotypes, at high $F$ values.

We therefore examined the fitness landscape directly, with the goal of characterizing the site multiplicity of all fit genotypes. First, we analytically estimated the frequency of genotypes with exactly $k$ binding sites at relative affinity $\geq 0.25$ (Materials and Methods), shown in figure 3*A*. We see that for the most part, genotypes with fewer sites are more abundant. Next, for each $k$, we sampled genotypes with exactly $k$ binding sites (uniformly at random), computed the fitness functional for each genotype and thus estimated the probability that such genotypes are fit ($F \geq 0.8$) (fig. 3*B*). Finally, multiplying the quantities shown in figure 3*A* and *B*, we obtained the relative proportion of $k$-site genotypes in the space of fit genotypes (fig. 3*C*): the three most abundant site-multiplicity values are $k = 6, 5, 7$, in that order, together accounting for about 90% of the total frequency. Thus, complex genotypes are indeed more common among all fit genotypes, and this explains their dominance in the results of evolutionary simulation. We also noted clear examples of how one genotype class can be evolutionarily preferred over another class (e.g., 5-site vs. 7-site genotypes, fig. 2*B*) due to greater frequency (fig. 3*C*), despite being less fit on average (fig. 3*B*).

### *Importance of Weak Binding Sites*

We next analyzed why complex sequences are frequent among fit genotypes. We hypothesized that the strength of binding sites play a major role here that complex genotypes make use of contributions from many suboptimal sites to achieve the same net occupancy of the TF on the enhancer as might be achieved through fewer closer-to-optimal sites. If this is the case, the complex genotype bias in the fitness landscape (fig. 3*C*) should become less prominent as we make the threshold for counting sites more stringent. We found this to be the case indeed, as shown in figure 3*D*. For instance, at the high threshold of relative affinity $= 0.8$, where only the optimal site gets counted, the mode of the observed distribution (site multiplicity $k = 2$) also corresponds to the most parsimonious (simplest) genotype(s) observed to achieve the fitness criterion of $F \geq 0.8$; in other words, the complex genotype bias is not seen. A direct examination of site strengths revealed that complex genotypes have weaker sites on average than simpler genotypes (fig. 3*E*).

Thus, broadly speaking, there are two types of fit genotypes: simple sequences, with few strong sites, and complex sequences, with more weak sites. Both types of genotypes can achieve high fitness, but complex sequences

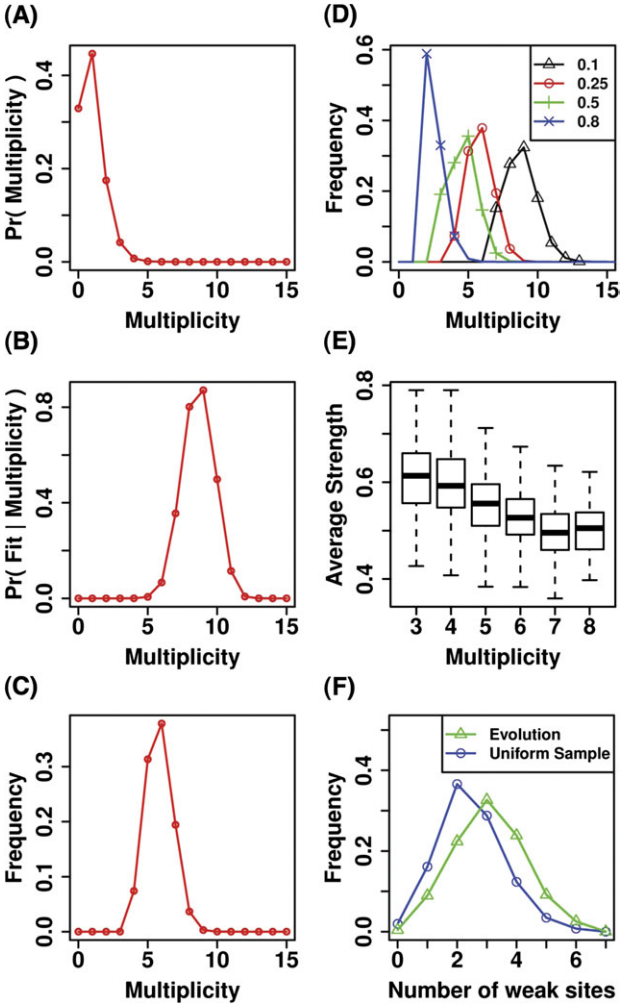


**Fig. 3.** Genotype frequency and properties of fit genotypes. (*A*) Relative frequency of genotypes with different site multiplicities: the number of sequences with $k$ binding sites (at relative affinity $\geq 0.25$) is estimated analytically, for each value of $k$. (*B*) Probability of a genotype with $k$ sites being fit ($F \geq 0.8$). (*C*) Frequency of $k$-site genotypes among all fit sequences, calculated by multiplying the relative frequency in (*A*) and the probability of being fit (*B*). (*D*) Same as plot (*C*), at different relative affinity thresholds (0.1, 0.25, 0.5, 0.8). (*E*) Average relative affinity of binding sites in $k$-site genotypes. (*F*) Histograms of subsites (relative affinity $< 0.25$), for evolutionary (green) and uniform (blue) samples of genotypes with $k = 6$ sites at relative affinity 0.25.

with weak sites are common in the evolutionary samples (fig. 2*B* and *D*) due to their high genotype frequency (fig. 3*C*). To illustrate the intuition behind this, we present a simple theoretical calculation. Let us characterize a genotype by the two integers $(k,m)$, where $k$ is the number of sites and $m$ is the strength of each site (defined here, for simplicity, by the number of mismatches relative to the optimal site). The abundance of $(k,m)$ genotypes can be calculated as

$$N(k,m) = \binom{L}{k} 2^k \left[ \binom{l}{m} 3^m \right]^k 4^{L-kl}, \qquad (1)$$

where $L$ is the length of the enhancer and $l$ is the length of each site. Using this formula and the parameter values in our setting ($L = 500$ and $l = 7$), we find that complex sequences

can be more common than simple ones (supplementary fig. S1, Supplementary Material online). For instance, we see that $N(2,1)$ is about 13 times larger than $N(1,0)$, that is, genotypes with two suboptimal sites are 13 times more frequent than genotypes with one optimal site. Similarly, $N(3,1)$ is ~6 times larger than $N(1,0)$. If we assume, for instance, that that one optimal site can be functionally replaced by a few suboptimal sites (e.g., 2–3 sites with 1 mismatch each), the class of fit genotypes will have a relative abundance of complex genotypes. This simplistic calculation, which is not tied to the precise genotype–phenotype mapping and its parameters, reveals the main idea behind an evolutionary origin of homotypic site clustering. In the supplementary text (Supplementary Material online), we explore a different theoretical model of binding sites where certain positions of a binding site, strong or weak, must remain invariant, and we find the same intuitive explanation of HTC to be revealed by this alternative model.

### *An Evolutionary Signature*

We designate the samples we obtained for studying the properties of the fit genotypes (fig. 3C) as "uniform samples" to distinguish them from evolutionary samples because the way evolution explores the fitness landscape depends on history (thus not uniform sampling). We noted that the evolutionary samples (fig. 2B) and uniform samples (fig. 3C) of the same population (i.e., all fit genotypes) have similar site multiplicity distributions, with most of their probability mass concentrated on the same values ($k = 5, 6$). However, the two distributions also have significant differences, for example, evolutionary samples include a greater representation of $k = 4$ genotypes compared with uniform samples (probability 0.21 vs. 0.07). This particular statistical observation led us to an interesting characterization of evolutionarily sampled genotypes. We first noted that the average fitness of $k = 4$ genotypes is comparable to that of $k = 5$ genotypes in the evolutionary samples (fig. 2C) but substantially lower in uniform samples (supplementary fig. S2, Supplementary Material online). In other words, evolution finds only the fittest ($F \approx 0.97$) among all fit $k = 4$ genotypes. Investigating this further, we noted that the $k = 4$ genotypes that evolution finds have unusually many "subsites" (sites below the strength threshold used), in addition to the 4 sites, that contribute to the occupancy and hence to fitness of the enhancer. This is clear from figure 3F, where histograms of subsite multiplicity show that fit genotypes sampled by evolution (green) and are significantly enriched in subsites compared with uniform samples (blue). In other words, evolutionary samples have a greater spread of site strengths than random expectation (represented by the uniform samples). This is what leads, in this case, to $k = 4$ genotypes found by evolution having unusually high fitness, and consequently, a higher relative frequency. Interestingly, this evolutionary signature was reported previously, as an abundance of subsites near functional binding sites, in a systematic analysis of 11 mammalian genomes (Reid 2007).

## Temporal Profile of Site Multiplicity in an Evolving Enhancer

We noted above (fig. 2) that the evolutionary process frequently samples complex genotypes after adaptation has been reached. However, it is plausible that parsimonious genotypes serve as the entry points to the space of fit genotypes that evolution explores. That is, evolution may be "stumbling into" parsimonious genotypes first because they have few sites, and after one fit genotype has been found, subsequent gain and loss of sites leads to more complex genotypes. We therefore asked if the evolving enhancer is parsimonious at the time of reaching adaptation and acquires additional sites postadaptation. Surprisingly, we found this not to be the case. Instead, we observed that the most parsimonious genotype reached by evolution (over a long period) is typically reached postadaptation. Figure 4A–C shows three typical simulations, in terms of how the site multiplicity changes with time, before as well as after adaptation. Note that in each simulation, the most parsimonious fit genotype (arrows) is encountered well after adaptation was reached (shown as the point of transition from gray to bold lines and by the presence of markers). This is true of most of our simulations: in 70% of our simulations, the most parsimonious fit genotype had at least two fewer sites than the genotype at which adaptation was reached (fig. 4D).

## Site Multiplicity Distributions in *Drosophila* Enhancers

Our study is based on the motif and concentration profile of the *Bicoid* TF, which activates expression in the anterior half of the blastoderm-stage embryo in *Drosophila*. Therefore, it is instructive to examine if our observations about site multiplicity distributions (in synthetic genotypes) are mirrored in real enhancers as well. We collected 21 *bona fide* enhancers that use *Bicoid* binding sites to drive anterior expression in the early embryo in *D. melanogaster*. For each enhancer, we also collected orthologs from (up to) five other moderately diverged species from the *Drosophila* group and computed their site multiplicity at the same threshold as in figure 2B above. These are shown in figure 5A, grouped by orthology. If one assumes that orthologous enhancers have similar fitness, this plot suggests that variability in site multiplicity and abundance of nonparsimonious genotypes is true of real fitness landscapes. However, orthologous enhancers may vary in their transcriptional outputs, and even if they have the same output, different orthologs may utilize *Bicoid* binding to different extents (for instance, by making use of other TFs). Therefore, we next estimated the occupancy of every enhancer in our collection (using the same procedure as in fig. 2E) and examined the site multiplicity distributions of enhancers grouped by occupancy. (Recall that in our simulations, postadaptation genotypes exhibit a narrow range of occupancy centered at ~7 [fig. 2E].) The results (fig. 5B) suggest that if we use estimated occupancy as a surrogate for the transcriptional effect of *Bicoid*, evolutionary samples of the fitness landscape exhibit (qualitatively) the same kind of variability of site counts that our theoretical study anticipates. For example, enhancers with estimated occupancy ~7 have site multiplicity in the range 2–5, with the median at 4. (Compare this with artificial evolutionary samples in fig. 2B, mostly

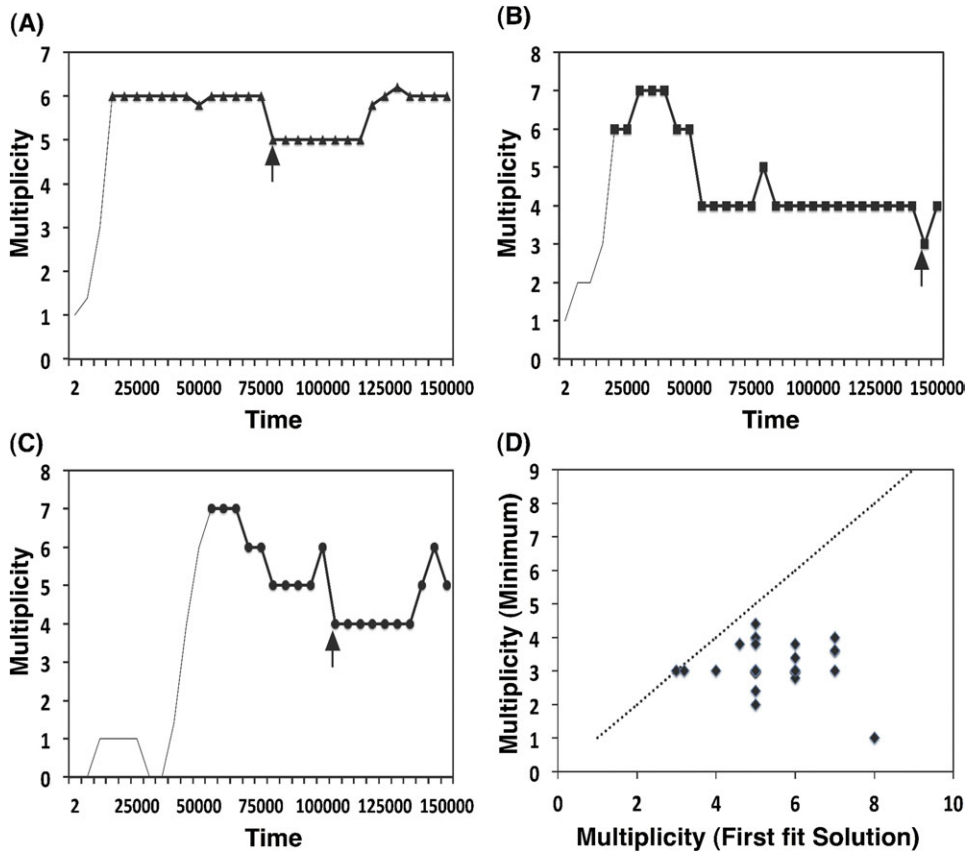


FIG. 4. Temporal dynamics of site multiplicity. (*A*–*C*) Average site multiplicity of genotypes in the evolving population, as a function of time, for three typical simulations. The gray part of each curve indicates postadaptation profile ($F \geq 0.8$); the black part indicates preadaptation. (*D*) Average site multiplicity of genotypes at adaptation (*x* axis) versus the minimum multiplicity encountered postadaptation (*y* axis). Each point represents one simulation.

with 3–7 sites and a median of 5.) That this is a qualitative rather than quantitative agreement is expected, since the quantitative model used in our simulations almost certainly misses certain aspects of the real enhancers' regulation.

## Other Potential Causes of Complex Genotype Bias

Finally, we investigated additional factors that may influence the emergence of a complex genotype bias in enhancers, including nonequilibrium sampling of the fitness landscape, short local duplications in DNA, and differences in stochasticity of gene expression induced by different genotypes.

### *Local Topography of Fitness Landscape*

The distribution of evolutionarily sampled genotypes (fig. 2) may depend on local properties of the fitness landscape. For instance, high fitness genotypes in a relatively "rugged" region may be sampled more or less frequently than similar-fitness genotypes in a smoother region (Weinberger 1991; Smith et al. 2002). We quantified the local ruggedness of the fitness landscape around a genotype by its "average correlation length" (ACL [Hordijk 1995]; supplementary fig. S9, Supplementary Material online) and found that genotypes with $k = 4$–8 sites had similar ACL, suggesting that topographical differences, at least to the extent characterized by the ACL score, do not significantly influence the complex genotype bias.

### *Effect of Local Duplications*

A remarkably high coverage of short tandem repeats has been observed in *Drosophila* enhancers (Sinha and Siggia 2005), suggesting that short local duplications may play an important role in regulatory sequence evolution and perhaps lead to homotypic site clustering. To investigate this, we compared the results of evolutionary simulations with substitutions, short insertions and deletions, to simulations where all or part of the insertions were local duplications. However, we observed no difference in either the complex genotype bias or the adaptation time in simulations with or without local duplications (data not shown). Future work will have to examine the role of local duplications in enhancer evolution under varying assumptions about the underlying indel and duplication rates and length distributions.

### *Noise Characteristics of Complex Genotypes*

The binding site composition of an enhancer has the potential to affect intrinsic noise (stochasticity) in gene expression levels and thus the robustness of biological processes (Raser and O'Shea 2004; Kaern et al. 2005). In particular, a recent study (Holloway et al. 2011) shows that

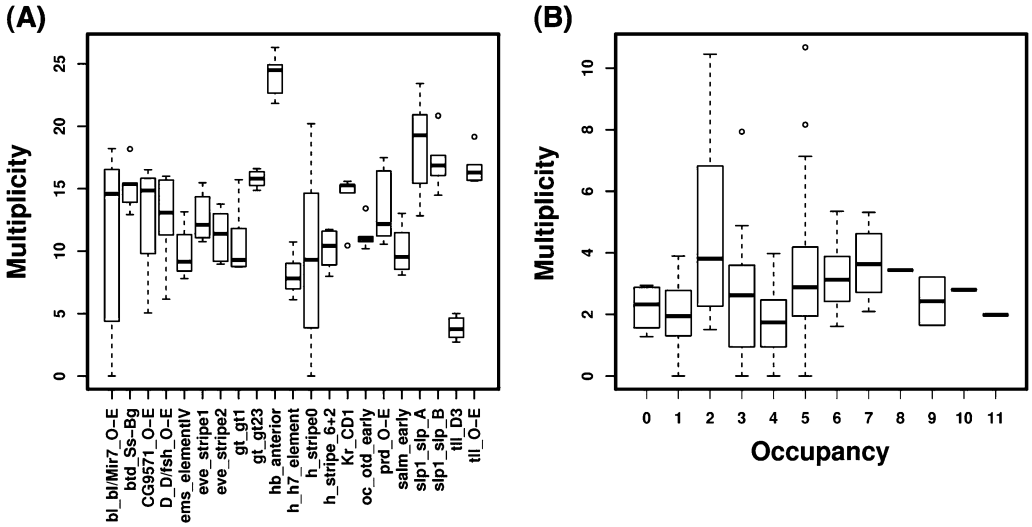


**Fig. 5.** Multiplicity and occupancy of orthologous enhancers of *Drosophila*. (*A*) Box plot of site multiplicities (at relative affinity $\geq 0.25$) for orthologs of *Bicoid*-driven enhancers. Multiplicity values are not normalized for length, since each orthology group has relatively little length variation. (*B*) Box plot of site multiplicities (*y* axis) for *Bicoid*-driven enhancers grouped by TF occupancy (*x* axis). Occupancy and multiplicity values are normalized by length, since each value of occupancy includes enhancers of widely different lengths.

greater number or strengths of *Bicoid* sites in the *hunchback* gene promoter leads to reduced noise in *hunchback* expression (while increasing the expression levels). We therefore asked if the (high fitness) genotypes sampled by our evolutionary simulations might reveal a correlation between site multiplicity and noise in gene expression. We estimated variance in TF occupancy on each enhancer (occupancy and expression level are correlated in our model) and found a strong negative correlation with site multiplicity (supplementary fig. S10*A*, Supplementary Material online). Importantly, this correlation exists despite the mean occupancy being roughly constant (supplementary fig. S10*B*, Supplementary Material online). Phenotypic consequences of reduced noise in expression may therefore be an important factor leading to the complex genotype bias observed in real enhancers. However, since such consequences were not factored into our fitness function, we conclude that the bias toward HTC can arise even in the absence of a noise–fitness relationship.

## Discussion

A fundamental aspect of understanding the complexity and design of a biological system is whether these features are functional requirements or consequences of the evolutionary process (Lynch 2007). For instance, a complex design may be chosen by evolution not because of any inherent functional advantages over alternative designs, but because it is more easily found by evolution (Soyer and Bonhoeffer 2006). We studied this question in the context of *cis*-regulatory sequences. The design feature we investigated is the HTC of TF binding sites, found in regulatory sequences across major animal kingdoms (Lifanov et al. 2003; Gotea et al. 2010). The relative simplicity of the system we studied, where the phenotype (expression pattern) of a sequence can be defined using a well-studied biophysical model, allows us to simulate its evolution and perform controlled analysis. Our results show that even when simpler designs exist for the desired expression pattern, relatively complex designs (genotypes with more sites) are more readily reached by evolution (fig. 2*B* and *D*). This is, to a large extent, because those complex sequences occupy a larger proportion of the space of fit genotypes (fig. 3*C*). There are more ways to "build" a fit enhancer with many weak sites than with a few strong sites, and this is why evolution finds the former type more often. We also observed a subtle but clear evolutionary signature in the synthetic enhancers: evolutionary samples tend to have a broader spread of site strengths (fig. 3*F*) than expected from a uniform sampling of all fit genotypes. We explored the temporal profiles of site multiplicity in an evolving enhancer, and found, somewhat surprisingly, that simpler designs are not necessarily the precursors of more complex designs that evolve postadaptation (fig. 4). We examined site multiplicities of *Bicoid*-driven enhancers in *Drosophila* species and found a characteristically broad range of multiplicities among enhancers grouped by orthology or by estimated *Bicoid* occupancy (fig. 5), providing empirical evidence for the complex genotype bias we observe in simulations. Finally, we investigated alternative sources of this bias and found that local topography of the fitness landscape (around a fit genotype) does not play a significant role nor does the phenomenon of short local duplications in the sequence, at least within the parameter ranges we explored. On the other hand, the higher fidelity (reduced noise in gene expression) associated with complex genotypes is a potential cause of their relative abundance, even though we did not explicitly demonstrate this within our simulation framework.

We note that to an extent, HTC does arise from functional requirements—if an enhancer driving the appropriate expression level requires an occupancy of say 5, it must harbor at least five sites; this is a functional constraint. At

the same time, it is accepted that multiple weak sites may function as well as one or few strong binding sites (Roider et al. 2007; Shultzaberger et al. 2010), suggesting that the neutral space (Wagner 2007) of fit genotypes may be highly diverse. We propose that this diversity is a key determinant of enhancer composition and that the required TF occupancy is more likely to be implemented through a greater number of sites (including suboptimal ones) than with the minimal number of optimal sites.

Earlier work has proposed specific mechanistic explanations of HTC that multiple sites may facilitate TF-DNA interaction synergistically (Giniger and Ptashne 1988; Lin et al. 1990; Anderson and Freytag 1991; Hertel et al. 1997; He et al. 2010) or that HTC can make sequences more robust to genetic and environmental perturbations (Ludwig et al. 1998), among others (Gotea et al. 2010). However, our simulations clearly showed a complex genotype bias even in the absence of cooperative interactions between sites (also see supplementary fig. S3, Supplementary Material online) and despite the fact that our fitness function does not incorporate robustness. Thus, we offer a plausible explanation for HTC that relies upon fairly general assumptions about the underlying biochemical model and fitness function. This provides a baseline that more specific mechanistic explanations may be compared with or used in conjunction with. We do note that our results rely upon contributions of multiple sites being free from spatial constraints, unlike what is proposed in enhanceosomal models of enhancer function (Arnosti and Kulkarni 2005). Without this assumption, calculation of the abundance of genotypes may favor simple instead of complex sequences. Many studies to date have found the arrangement of binding sites in metazoan enhancers to be extremely flexible (Brown et al. 2007; He et al. 2009; Liberman and Stathopoulos 2009), supporting our assumption, but this issue is currently open to debate.

Our intuitive explanation of HTC is based on the assumption that the function of a strong binding site can be replaced by multiple weak ones. However, there are many reported cases where an enhancer may harbor multiple high-affinity binding sites. We hypothesize several possible explanations: for instance, some enhancers may demand a high level of TF affinity that requires multiple high-affinity sites; the enhanceosome model as explained above makes it impossible to trade one strong site for multiple weak ones (Crocker et al. 2010). Also, nonadaptive forces such as short tandem duplication (Sinha and Siggia 2005) may facilitate the occurrence of multiple high-affinity sites. A recent study (Paixao and Azevedo 2010) examines the multiplicity of binding sites in enhancers and uses simulations to show that this is largely due to recombination and weak direct selection for multiplicity. However, the definition of multiplicity (as the presence of two or more perfect binding sites) by Paixao et al. is very different from our definition, making its central question distinct from ours. Khatri et al. (2009) studied the evolution of enhancer sequences using a model system similar to ours but focused on the question of whether the optimal phenotype is reached (or not), in an adaptive process.

One way to interpret our results is that in genotypes found by evolution, the desired function (phenotype) is distributed into multiple weak components, instead of being concentrated on one or two strong ones. Such "distributed" designs, if allowed, may be a common feature of other systems. For example, signal transduction processes are often characterized by a long cascade of signaling events, where each step may serve only a small piece of the overall function of the pathway (e.g., extent of signal amplification) (Li and Qian 2003; Soyer and Bonhoeffer 2006). Our analysis suggests that a distributed design may in fact be a consequence of the evolutionary process, where both fitness and abundance of genotypes are important determining factors for the sampled designs.

The framework developed for this study can be used more broadly to explore the link between enhancer composition and evolution, for example, to explain binding site turnover rates in developmental enhancers that interact with multiple TFs (Kim et al. 2009) or to understand the effects of mechanistic features such as synergistic activation, DNA-binding cooperativity, and short-range repression (He et al. 2010) on evolutionary dynamics and sequence-level properties of enhancers. Two features of our framework—the ability to map any given sequence into its regulatory function (without using arbitrary rules on numbers of sites of various TFs or arbitrary thresholds to distinguish sites from nonsites) and an efficient implementation of the function model as well as the Wright–Fisher process—make it particularly suitable for such studies. Our computer program will be available at the Sinha lab website (upon publication).

## Supplementary Material

Supplementary text and figures S1–S12 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

This work was supported in part by the National Science Foundation (career grant number DBI-0746303) and in part by the National Institute of Health (grant number 5R01GM085233A). We thank Eric Siggia for comments on the manuscript during its preparation. We also thank Majid Kazemian and Md. Abul Hassan Samee for their help with data collection.

## References

Anderson GM, Freytag SO. 1991. Synergistic activation of a human promoter in vivo by transcription factor Sp1. *Mol Cell Biol.* 11(4):1935–1943.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J Cell Biochem.* 94(5):890–898.

Bergman CM, Carlson JW, Celniker SE. 2005. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* 21(8):1747–1749.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A.* 99(2):757–762.

Brown CD, Johnson DS, Sidow A. 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* 317(5844):1557–1560.

Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A.* 100(9): 5136–5141.

Coleman RA, Pugh BF. 1995. Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J Biol Chem.* 270(23):13850–13859.

Crocker J, Potter N, Erives A. 2010. Dynamic evolution of precise regulatory encodings creates the clustered site signature of enhancers. *Nat Commun.* 1:99.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* 148(4):1667–1686.

Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457(7226):215–218.

Giniger E, Ptashne M. 1988. Cooperative DNA binding of the yeast transcriptional activator GAL4. *Proc Natl Acad Sci U S A.* 85(2):382–386.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* 20(5):565–577.

Halfon MS, Gallo SM, Bergman CM. 2008. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res.* 36(Database issue):D594–D598.

He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH, Zhong S. 2009. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One* 4(12):e8155.

He X, Samee MA, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol.* 6(9):e1000935.

Hertel KJ, Lynch KW, Maniatis T. 1997. Common themes in the function of transcription and splicing enhancers. *Curr Opin Cell Biol.* 9(3):350–357.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. *Genetics* 177(3):1725–1731.

Holloway DM, Lopes FJ, da Fontoura Costa L, Travencolo BA, Golyandina N, Usevich K, Spirov AV. 2011. Gene expression noise in spatial patterning: hunchback promoter structure affects noise amplitude and distribution in Drosophila segmentation. *PLoS Comput Biol.* 7(2):e1001069.

Hordijk W. 1995. A measure of landscapes. *Evol Comput.* 4:335–360.

Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet.* 6(6):451–464.

Kazemian M, Blatti C, Richards A, et al. 2010. Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. *PLoS Biol.* 8(8):e1000456.

Khatri BS, McLeish TC, Sear RP. 2009. Statistical mechanics of convergent evolution in spatial patterning. *Proc Natl Acad Sci U S A.* 106(24):9564–9569.

Kim J, He X, Sinha S. 2009. Evolution of regulatory sequences in 12 Drosophila species. *PLoS Genet.* 5(1):e1000330.

Kim JG, Takeda Y, Matthews BW, Anderson WF. 1987. Kinetic studies on Cro repressor-operator DNA interaction. *J Mol Biol.* 196(1):149–158.

Li G, Qian H. 2003. Sensitivity and specificity amplification in signal transduction. *Cell Biochem Biophys.* 39(1):45–59.

Li L, Zhu Q, He X, Sinha S, Halfon MS. 2007. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. *Genome Biol.* 8(6):R101.

Liberman LM, Stathopoulos A. 2009. Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. *Dev Biol.* 327(2):578–589.

Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. 2003. Homotypic regulatory clusters in Drosophila. *Genome Res.* 13(4):579–588.

Lin YS, Carey M, Ptashne M, Green MR. 1990. How different eukaryotic transcriptional activators can cooperate promiscuously. *Nature* 345(6273):359–361.

Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* 125(5):949–958.

Lusk RW, Eisen MB. 2010. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genet.* 6(1):e1000829.

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A.* 104(Suppl 1): 8597–8604.

Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A.* 99(2):763–768.

Mustonen V, Kinney J, Callan CG Jr, Lassig M. 2008. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci U S A.* 105(34):12376–12381.

Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, Papatsenko D, Small S. 2005. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc Natl Acad Sci U S A.* 102(14):4960–4965.

Paixao T, Azevedo RB. 2010. Redundancy and the evolution of cis-regulatory element multiplicity. *PLoS Comput Biol.* 6(7): e1000848.

Porcher A, Dostatni N. 2010. The bicoid morphogen system. *Curr Biol.* 20(5):R249–R254.

Raser JM, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. *Science* 304(5678):1811–1814.

Reid ID. 2007. Transcription factor binding site turnover in mammals. Montreal (Canada): McGill University.

Roider HG, Kanhere A, Manke T, Vingron M. 2007. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23(2):134–141.

Sauer F, Hansen SK, Tjian R. 1995. DNA template and activator-coactivator requirements for transcriptional synergism by Drosophila bicoid. *Science* 270(5243):1825–1828.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 451(7178):535–540.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A.* 102(27): 9541–9546.

Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol.* 181(2):211–230.

Shultzaberger RK, Malashock DS, Kirsch JF, Eisen MB. 2010. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. *PLoS Genet.* 6(7): e1001042.

Sinha S, Adler AS, Field Y, Chang HY, Segal E. 2008. Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res.* 18(3):477–488.

Sinha S, Siggia ED. 2005. Sequence turnover and tandem repeats in cis-regulatory modules in Drosophila. *Mol Biol Evol.* 22(4): 874–885.

Smith T, Husbands P, Layzell P. 2002. Fitness landscapes and evolvability. *Evol Comput.* 10(1):1–34.

Soyer OS, Bonhoeffer S. 2006. Evolution of complexity in signaling pathways. *Proc Natl Acad Sci U S A.* 103(44):16337–16342.

Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16(1):16–23.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genetics* 172(3): 1607–1619.

Wagner A. 2007. Robustness and evolvability in living systems. Princeton (NJ): Princeton University Press.

Weinberger ED. 1991. Local properties of Kauffman's N-k model: a tunably rugged energy landscape. *Phys Rev A.* 44:6399–6413.

Zinzen RP, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr Biol.* 16(13):1358–1365.